Realisierung von Missing-Data-Ersetzungstechniken innerhalb statistischer Programmpakete und ihre Leistungsfähigkeit

Schnell, Rainer

Veröffentlichungsversion / Published Version
Sammelwerksbeitrag / collection article

Realisierung von Missing-Data-Ersetzungstechniken innerhalb statistischer Programmpakete und ihre Leistungsfähigkeit

von Rainer Schnell

1. Einleitung

In den meisten Datensätzen der empirischen Sozialforschung fehlen einzelne Angaben für Untersuchungsobjekte. Das Ausmaß fehlender Werte (missing data: MD) variiert abhängig vom Untersuchungsgegenstand, der Untersuchungspopulation, der Datenerhebungsmethode, der Datenbereinigung usw. In den meisten Fällen liegt der Prozentsatz der MD zwischen 1% und 10%. Trotz dieser relativ großen Anteile werden in den meisten Analysen die fehlende Werte einfach ignoriert. Santos (1981:22) hat diese Methode als die "Do Nothing Technique" bezeichnet. Um unverzerrte Ergebnisse zu erhalten, muß man bei Anwendung dieser Methode annehmen, daß die Daten zufällig fehlen ("missing at random": MAR).

Obwohl exakte mathematische Definitionen von MAR vorliegen (Rubin 1976, Little/Rubin 1983:216, Anderson/Basilevsky/Hum 1983:416-417), reicht für die meisten praktischen Zwecke die einfache Definition von Glasser (1964:834) aus: Es sei $B_{i'j}$ eine (n*p)-Indikatormatrix, wobei $B_{i'j} = 0$ falls $x_{i'j}$ fehlt und $B_{i'j} = 1$ andernfalls. Falls

$$E(B_{i'1},B_{i'2},...B_{i'p}) = E(B_{i'1})\,E(B_{i'2})\,...\,E(B_{i'p})$$

dann werden die Daten als zufällig fehlend betrachtet. MAR in diesem Sinn ist die stochastische Unabhängigkeit des Fehlens einer Variablen vom Fehlen jeder anderen Variablen im Datensatz.

Falls die MAR-Annahme korrekt ist, dann ist die Beschränkung der Analyse auf diejenigen Fälle, für die alle Variablen gültige Werte besitzen ("listwise deletion") nur ein weiterer Schritt bei der Ziehung einer Zufallsstichprobe. In diesem Fall sind die Schätzungen unverzerrt, lediglich die Mächtigkeit der statistischen Tests ist (auf Grund der kleineren Stichprobe) geringer (vgl. Little/Rubin 1987). Um die MAR-Annahme zu testen stehen eine Reihe von Verfahren zu Verfügung (Cohen/Cohen 1975, Frane 1978, Hill/Dixon 1981, Möntmann/Bollinger/Herrman 1983). Obwohl diese prinzipiell nicht in der Lage sind, jeden möglichen systematischen Prozeß zu entdecken der fehlende Werte hervorbringen kann ("missing data generating

process", vgl. hierzu Rubin 1976), reichen in der Praxis meist die einfachsten Testverfahren schon aus, um jede Hoffnung auf zufällig fehlende Werte zu zerstören (Frane 1978:27). Bemerkenswerterweise existieren nur sehr wenige Ansätze zur Analyse von Datensätzen mit systematisch fehlenden Werten, die dann aber auch immer nur in sehr speziellen Fällen angewendet werden können (vgl. Little/Rubin 1987:218-243). Wenn MD-Probleme in der Praxis überhaupt angegangen werden, dann werden nahezu ausschließlich sehr einfache Schätz- und Ersetzungsverfahren ("imputation methods") verwendet.

2. Missing-Data-Techniken

Da es hier unmöglich ist, einen vollständigen Überblick über die Literatur zu MD-Techniken zu geben [1], sollen nur die wichtigsten Methoden für metrische Daten [2] erwähnt werden.

Grob kann zwischen Methoden zur Schätzung von Parametern auf der Basis von Datensätzen mit fehlenden Werten und Verfahren zur Ersetzung fehlender Werte durch ihre Schätzung ("imputation methods") unterschieden werden.

Die allgemein bekanntesten Verfahren der ersten Gruppe sind die verschiedenen "missing data correlations", z.B. "pairwise", "listwise" or "allvalue". Zu den Parameterschätzverfahren gehören aber auch zwei vergleichsweise neuere Verfahren: Der sogenannte "EM-Algorithmus" und eine spezielle Klasse ökonometrischer Modelle, die "sample selection-" bzw. "Nelson-Heckman"-Modelle (Heckman 1976, Nelson 1977).

Die "sample selection"-Modelle (Berk/Ray 1982; Berk 1983; Little 1983) erlauben unverzerrte Schätzungen von Regressionsparametern trotz systematisch fehlender Werte. Diese Modelle basieren alle auf der Annahme eines fehlende Werte generierenden Prozesses, bei dem ein Wert dann fehlt, wenn eine latente Variable einen unbeobachteten stochastischen Schwellenwert überschreitet. Neben einigen Verteilungsannahmen ist das Hauptproblem [3] dieser Modelle die Spezifikation eines theoretischen Modells des "Missing

[1] Eine knappe Einführung geben Anderson/Basilevsky/Hum (1983). Eine ausführliche Darstellung dieser und anderer, eher selten verwendeter Methoden gibt Schnell (1986:83-137).

[2] Für Methoden zur Behandlung (auch systematisch) fehlender Werte bei qualitativen Daten vgl. vor allem Fuchs (1982), Fay (1986), Little/Rubin (1987) und Haagenaars (1988).

[3] Die Berechnung solcher Modelle ist mit speziellen Programmen sehr einfach geworden. Ein solches auch für IBM-PC's verfügbares Programm ist LIMDEP (Greene 1986).

Data generierenden Prozesses". Unglücklicherweise existieren solche Modelle außerhalb der Ökonometrie in den Sozialwissenschaften kaum, so daß nur selten Anwendungen gefunden werden können.

Die neuere MD-Literatur wird durch den "EM-Algorithmus" als Parameterschätzverfahren dominiert. Der "expectation maximization" (EM)-Algorithmus ist eine allgemeine Methode [4] zur Berechnung von Maximum-Likelihood-Schätzern, die (im Prinzip) für jedes MD-Problem verwendet werden könnte (Orchard/Woodbury 1972; Dempster/Laird/Rubin 1977). Der EM-Algorithmus ist eine Iterationsmethode, bei der jede Iteration aus einem E-Schritt, in dem die Log-Likelihood-Funktion berechnet wird, und einem M-Schritt, in dem diejenigen Parameter berechnet werden, die die Log-Likelihood-Funktion maximieren, besteht (vgl. Little 1983:369-371). In diesem Zusammenhang kann der EM-Algorithmus folgendermaßen dargestellt werden:

1. Berechnung eines Ausgangsmittelwertvektors und einer Ausgangskovarianzmatrix (listwise);
2. Für jeden Fall mit MD: Partitionieren des Mittelwertvektors und der Kovarianzmatrix in Matrizen für die vollständigen und unvollständigen Variablen;
3. Schätzung der MD durch multiple Regression, wobei alle vorhanden Variablen dieses Falles für die Schätzung verwendet werden.
4. Berechnung eines neuen Mittelwertvektors und einer neuen Kovarianzmatrix;
5. Korrektur der Kovarianzmatrix: Für jeden Fall mit MD wird die residuale Kovarianz der Prädiktorvariablen zum entsprechenden Element der Kovarianzmatrix addiert;
6. Berechnung eines Konvergenzkriteriums; falls keine Konvergenz: Wiederholung der Schritte 2-6; sonst Stopp.

Da diese Form des EM-Algorithmus also auch die fehlenden Werte ersetzt, gehört es zu beiden Gruppen von Verfahren. Eine Implementation dieser Methode (mit der Ausnahme von ALLVALUE in Schritt 1) findet sich in BMDPAM mit der "ML"-Option im "ESTIMATE TYPE="-Kommando.

Natürlich ist es möglich fehlende Werte direkt mit anderen Methoden zu schätzen. Die Ersetzung fehlender Werte durch Schätzungen wird "imputation" genannt. Imputation besitzt zwei Vorteile:

1. Nach der einmaligen Ersetzung liegt ein vollständiger "Allzweck"-Datensatz vor, der mit traditionellen Methoden (und Programmen) für vollständige Daten analysiert werden kann;

[4] Die Bezeichnung als 'Algorithmus' (durch Dempster/Laird/Rubin 1977) ist daher äußerst irreführend.

2. Durch die Imputation kann möglicherweise die Verzerrung durch systematisch fehlende Werte ausgeglichen werden, vor allem dann, wenn die Ersetzung durch die Produzenten des Datensatzes vorgenommen wird, die ihre möglicherweise vorhandenen zusätzlichen Informationen über den MD generierenden Prozeß bei der Schätzung verwenden können.

Eine nützliche Unterscheidung zwischen "informativen" und "nicht informativen" Imputationsmethoden stammt von Santos (1981). Informative Imputationen wurden von Santos (1981:30) als Methoden definiert, die zusätzliche Informationen bei der Zuweisung von Werten zu unvollständigen Fällen verwenden.

Die einfachste uninformative Methode ist das pure Raten eines fehlenden Wertes, die nächst einfache Methode ist die Ersetzung eines MD mit dem Mittelwert der vorhanden Werte (im Folgenden: MEAN). Bei der Schätzung von Korrelationen wird diese Methode häufig als "Wilk's Methode" bezeichnet. Eine Reihe von Statistikprogrammpaketen (z.B. SPSS-X) bieten diese Methode als "mean substitution" bei der Schätzung der Kovarianzmatrix für Regressionen an.

Ein sehr einfaches Beispiel einer informativen Ersetzung ist die "Zellenmittelwertersetzung" (im Folgenden: CELL). Hier werden die MD durch den Mittelwert einer Subgruppe ersetzt, zu der der unvollständige Fall gehört. Das Problem dieser und anderer subgruppenbasierter Techniken liegt in der Bestimmung der Subgruppen, die in diesem Zusammenhang als "Imputationsklassen" bezeichnet werden. Imputationsklassen sollten nur sehr geringe Binnengruppenvarianzen besitzen, aber soviel Zwischengruppenvarianz wie möglich. Weiterhin sollte die Zahl der Klassen gering und die Klassifikationsvariablen weitgehend unabhängig voneinander sein (Rizvi 1983:306). Damit eine Variable überhaupt als Klassifikationsvariable verwendet werden kann, muß der Anteil fehlender Werte für diese Variable sehr klein sein. Häufig wird eine Form der bekannten "Automatic Interaction Detector (AID)"-Technik [5] (oder eine andere Clusteranalysetechnik [6]) verwendet um die Klassifikationsvariablen zu bestimmen [7].

[5] Die klassische Arbeit ist Sonquist/Baker/Morgan (1971). Dillon/Goldstein (1984:183-186) geben eine kurze Einführung. Die beste technische Dokumentation von AID findet sich bei Hartigan (1975). Eine allgemeine Kritik von AID gibt Fielding (1979).

[6] Eine ausführliche Behandlung von MD-Problemen bei Clusteranalysen findet sich in zwei exzellenten Arbeiten von Wishart (1978, 1985).

[7] Vgl. Rizvi (1983). Für AID-Techniken in diesem Zusammenhang, vgl. Kalton/Santos (1983:84-87, 93-96) und Chapman (1983). Vgl. ferner Oh/Scheuren (1983:157-158). Für eine Kritik der theoretischen Annahmen, vgl. Schnell (1986:104-106).

Eine etwas interessantere Menge von Techniken bilden die sogenannten "Hot-Deck"-Methoden, die vor allem in der amtlichen Statistik, insbesondere bei Volkszählungen, verwendet werden. Obwohl es keine einheitliche Definition von Hot-Decks gibt (Ford 1983:185-186), ist allen Definitionen die Erwähnung des stochastischen Charakters, die Verwendung von Daten des unvollständigen Datensatzes und des "Dopplungscharakters" des Verfahrens gemein. Dies wird klarer, wenn wir zwischen zwei Arten von Hot-Deck-Methoden unterscheiden: Sequentiellen und simultanen Hot-Decks.

Sequentielle Hot-Decks (auch als "traditionelle Hot-Decks" bezeichnet, vgl. Kalton/Kasprzyk 1982:23) beginnen mit der Bestimmung der Imputationsklassen und der Startwerte für die fehlenden Variablen jeder Imputationsklasse. Die Startwerte sind häufig die Zellenmittelwerte der Variablen in früheren Untersuchungen. Der unvollständige Datensatz wird dann sequentiell abgearbeitet. Für jeden Fall des unvollständige Datensatzes wird seine Imputationsklasse bestimmt. Trifft das Programm auf einen gültigen Wert des aktuellen Falles, so wird dieser Wert in die entsprechende Imputationsklasse kopiert (der Fall wird ein "Donor"). Trifft das Programm auf einen fehlenden Wert, so wird dieser durch den derzeitigen Inhalt der Imputationsklasse ersetzt (der Fall wird ein "Rezipient"). Das Verfahren wird bis zum Ende des Datensatzes fortgesetzt, anschließend besitzt jeder Fall für die vervollständigte Variable einen gültigen Wert. Falls mehr als eine Variable ersetzt wird, ist es möglich, daß bestehende Editierungsregeln verletzt werden, wenn die Daten eines bestimmten Falles von verschiedenen "Donors" stammen: So z.B. wenn einer der "Donors" weiblich und ein anderer "Donor" männlich ist und das Geschlecht des zweiten für die Imputation verwendet wird, aber die Anzahl der Schwangerschaften des ersten. In tatsächlichen Hot-Deck-Systemen sind die Editierungsregeln weit komplizierter als die des Beispiels. Die einfachste Lösung dieses Problems besteht darin, zunächst den vollständigen Datensatz gemäß den Editregeln zu bereinigen und dann den bereinigten vollständigen Datensatz so zu verwenden, daß die Daten jedes unvollständigen Falles für alle durch mögliche Editregeln betroffenen Variablen durch einen Vektor von Variablen eines vollständigen Falles ersetzt werden. Dabei werden auch vorhandene gültige Werte durch Werte des Donors ersetzt. Sequentielle Hot-Decks werfen eine Reihe weiterer Probleme auf, so z.B. die Bestimmung der Imputationsklassen, ungleichmäßige Benutzung von "Donoren" (d.h. manche Fälle werden extrem häufig zur Imputation herangezogen), und Abfolgeeffekte [8].

[8] Detaillierte und praxisbezogene Darstellungen dieser Probleme geben Sande (1982) und Ford (1983).

Die seltener verwendeten [9] simultanen Hot-Decks verwenden sogenannte "matching variables" (entsprechend den Klassifikationsvariablen), wobei aber jeder unvollständige Fall durch die Daten desjenigen vollständigen Falles vervollständigt wird, zu dem er die kleinste Distanz auf den Matchvariablen aufweist. Zumeist wird hierbei die euklidische Distanz

$$d_{i,j} = \left[\sum_{k=1}^{p} (x_{i,k} - x_{j,k})^2 \right]^{1/2}$$

verwendet, gelegentlich auch die Mahalanobis-Distanz:

$$d_{i,j} = (X_i - X_j)' S^{-1} (X: X_j),$$

wobei X_i ein (1*p)-Vektor der Variablen des Falles i, X_j ein (1*p)- Vektor der Variablen des Falles j, und S^{-1} die Inverse der (p*p)- Kovarianzmatrix der Variablen ist.

Sande (1979:248) betrachtet die Wahl der Distanzfunktion als unkritisch. Methoden mit solchen Ersetzungsregeln werden gelegentlich als "nearest-neighbor"-Methoden [10] bezeichnet.

Selbstverständlich lassen sich viele multivariate Verfahren zur Schätzung fehlender Werte verwenden. Eine der frühesten Methoden stammt von Dear (1959). Dear berechnet die erste Hauptkomponente einer über Mittelwerte vervollständigten Datenmatrix. Der zum größten Eigenwert der Korrelationsmatrix gehörende normalisierte Eigenvektor g wird zusammen mit der z-transformierten Datenmatrix ($z_{i,j} = O$, falls $x_{i,j}$ fehlt) zur Schätzung verwendet. Ein fehlender Wert $z_{i,j}$ wird geschätzt durch:

$$z_{i,j} = g_j \sum_{k=1}^{p} z_{i,k}\, g_k$$

und dann wieder in die ursprünglichen Einheiten transformiert [11].

[9] Selbst mit spezieller Software sind diese Techniken weit aufwendiger als traditionelle Hot-Decks. Bei großen Datensätzen (Mikrozensen, Volkszählungen etc.) dürften solche Verfahren unbezahlbar werden.

[10] Einzelheiten zu simultanen Hot-Decks finden sich bei Vacek/Ashikaga (1980) and Sande (1979).

[11] Diese Beschreibung basiert auf Hamilton (1975:27). Die Originalarbeit war mir nicht zugänglich.

Offensichtlich ist es möglich fehlende Werte durch Regressionsverfahren zu schätzen. Dies wurde schon 1960 von Buck vorgeschlagen. Buck's Methode beginnt mit der (p*p) Korrelationsmatrix der vollständigen Fälle. Für jeden Fall i und jede unvollständige Variable j ($j=1,2..m_i$) wird der fehlende Wert durch eine multiple Regression mit allen vollständigen ($p-m_i$) Variablen des Falls i als unabhängige Variablen geschätzt. Falls jede Kombination fehlender Variablen möglich ist und k Variablen fehlen, müssen maximal

$$p \binom{p-1}{k-1}$$

verschiedene Regressionsgleichungen berechnet werden (Buck 1960:303). Da jeder geschätzte fehlende Wert exakt auf der Regressionsgeraden liegt, wäre die resultierende Kovarianzmatrix des vervollständigten Datensatzes verzerrt. Buck (1960:304) schlägt daher eine Korrektur der Kovarianzmatrix (A) des vervollständigten Datensatzes vor. Falls die Variable j für einen Anteil I_j des Samples fehlt, wird die korrigierte Varianz $a_{j'j}$ berechnet als

$$a_{j'j} = a_{j'j} + I_j/c_{j'j}$$

wobei $c_{j'j}$ das Diagonalelement von A^{-1} ist. Diese Korrektur wird nicht immer in der Literatur erwähnt und jede Art von Regressionsmethode als "Buck's method" bezeichnet.

Obwohl eine große Zahl verschiedener Abwandlungen von Regressionstechniken vorgeschlagen wurde (vgl. Schnell 1986:117-119) werden diese aber nur selten verwendet. Lediglich iterierte Formen von multiplen Regressionen sind üblich. Eine iterierte multiple Regressionsmethode zur Schätzung fehlender Werte (ohne Korrektur der Kovarianzmatrix) wurde 1959 von Federspiel, Monroe & Greenberg vorgeschlagen. Versionen solcher Subroutinen finden sich z.B. in den speziellen MD-Programmen von Berger (1979) und O'Grady (1982). Schmee/Hahn (1979) schlugen eine "iterated least square"-Technik für die Lösung von Censoring-Problemen vor. Interessanterweise ist eine iterative Buck-Methode (mit Korrektur) mit dem oben erwähnten EM-Algorithmus identisch, wie Beale/Little (1975:134-137) gezeigt haben.

3. Realisierung mit Standardpaketen

Nur wenige der erwähnten Techniken sind als Prozeduren in den Standardstatistikprogrammpaketen wie SPSS-X oder SAS enthalten (im Gegensatz zu BMDP, welches mit dem Programm AM (Autor: James Frane) die

meisten MD-Ersetzungstechniken, z.B. MEAN, CELL, multiple Regression und iterierte multiple Regression zur Verfügung stellt). Trotzdem ist es möglich, die meisten Techniken innerhalb der Standardpakete mit deren Datentransformations- und Fallselektionsmöglichkeiten zu realisieren.

Im Folgenden sollen einige Möglichkeiten gezeigt werden, wie mit einem einfachen derzeitigen Statistiksystem (SPSS-PC+) Ersetzungen vorgenommen werden können. Mit nur kleinen Änderungen funktionieren diese Methoden auch mit SPSS-9 und SPSS-X. Andere Pakete (z.B. SAS, BMDP oder SYSTAT) erfordern etwas größere Änderungen, aber die Logik der Prozeduren bleibt erhalten. Selbstverständlich können auch die kompliziertesten Techniken mit Hochsprachen wie z.B. PASCAL leicht realisiert werden. Für GAUSS liegt das Programm MISS vor, das viele der erwähnten Techniken zur Verfügung stellt.

Die Beispiele basieren auf einem kleinen Datensatz (MISSING.SYS) mit 5 Variablen (V1 to V5) mit fehlenden Werten. Der Datensatz enthält darüber hinaus drei zusätzliche Variablen AUX1 bis AUX3 ohne fehlende Werte. Die letzte Variable im Datensatz ist eine vollständige Gruppierungsvariable GROUP. Es muß angenommen werden, daß zumindest ein Fall innerhalb jeder Gruppe (GROUP) vollständige Daten für alle Variablen besitzt.

Eine der einfachsten Imputationstechniken ist die Ersetzung durch den Mittelwert. Diese kann z.B. durch den folgenden SPSS-PC-Job erfolgen:

```
GET FILE='MISSING.SYS'.
COMPUTE DUMMY=1.
SAVE OUTFILE='SCRATCH'.
AGGREGATE OUTFILE= *
/BREAK=DUMMY
/M1 TO M5 = MEAN (V1 TO V5).
JOIN MATCH /TABLE=* /FILE='SCRATCH'/BY DUMMY.
IF (MISSING(V1)) V1=M1.
IF (MISSING(V2)) V2=M2.
IF (MISSING(V3)) V3=M3.
IF (MISSING(V4)) V4=M4.
IF (MISSING(V5)) V5=M5.
```

Die Variable DUMMY wird lediglich durch die Syntax des AGGREGATE-Kommandos erzwungen.

Eine Ersetzung durch Zellenmittelwerte (alle fehlenden Werte werden durch den Mittelwert einer Variablen in einer Zelle, zu der der Fall gehört, ersetzt) mit der Variablen GROUP als Gruppierungsvariable benötigt nur eine kleine Modifikation dieses Jobs:

```
GET FILE='MISSING.SYS'.
COMPUTE SEQUENCE=$CASENUM.
SORT CASES BY GROUP.
SAVE OUTFILE='SCRATCH'.
AGGREGATE OUTFILE= *
/PRESORTED
/BREAK=GROUP
/M1 TO M5 = MEAN (V1 TO V5).
JOIN MATCH /TABLE=* /FILE='SCRATCH'/BY GROUP.
IF (MISSING(V1)) V1=M1.
IF (MISSING(V2)) V2=M2.
IF (MISSING(V3)) V3=M3.
IF (MISSING(V4)) V4=M4.
IF (MISSING(V5)) V5=M5.
SORT CASES BY SEQUENCE.
```

Die erste Sortierung ist notwendig für das JOIN MATCH-Kommando, die letzte Sortierung stellt die ursprüngliche Abfolge der Fälle im Datensatz wieder her.

Hot-Deck-Verfahren können in vielen Varianten realisiert werden. Ein einfache Methode für eine variablenweise Ersetzung ist der folgende Job:

```
GET FILE='MISSING.SYS'.
COMPUTE SEQUENCE=$CASENUM.
COUNT NMISS=V1 TO V5 (MISSING).
SORT CASES BY GROUP NMISS.
IF (MISSING(V1)) V1=LAG(V1).
IF (MISSING(V2)) V2=LAG(V2).
IF (MISSING(V3)) V3=LAG(V3).
IF (MISSING(V4)) V4=LAG(V4).
IF (MISSING(V5)) V5=LAG(V5).
SORT CASES BY SEQUENCE.
```

Die Berechnung von NMISS und das Sortieren nach GROUP und NMISS garantieren, daß der erste Fall jeder Gruppe für jede Variable gültige Werte besitzt (falls die oben genannte Voraussetzung über die Struktur des Datensatzes erfüllt ist).

Diese Art des Hot-Decks kann bestehende Edit-Regeln verletzen, falls die zugeschrieben Werte von verschiedenen "Donors" stammen. Um dies zu vermeiden, führt eine leichte Änderung zu einer fallweisen Ersetzung der fehlenden Werte:

```
GET FILE='MISSING.SYS'.
COMPUTE SEQUENCE=$CASENUM.
COUNT NMISS=V1 TO V5 (MISSING).
SORT CASES BY GROUP NMISS.
IF (NMISS NE 0) V1=LAG(V1).
IF (NMISS NE 0) V2=LAG(V2).
IF (NMISS NE 0) V3=LAG(V3).
IF (NMISS NE 0) V4=LAG(V4).
IF (NMISS NE 0) V5=LAG(V5).
SORT CASES BY SEQUENCE.
```

Obwohl die Resultate zweier Simulationen (Kaiser 1983, Schnell 1986; abweichende Ergebnisse werden von Vacek/Ashikaga 1980 berichtet) darauf hinweisen, daß Verfahren mit "nearest-neighbor rule" zur Ersetzung zu stark verzerrten Kovarianzschätzungen führen, soll eine Version dieser Technik hier illustriert werden. Da keine SPSS-Version die Verwendung eines berechneten Resultates als Parameter eines Kommandos erlaubt, muß dies in zwei Schritten geschehen. Der erste Job berechnet die Z-Werte der zusätzlichen Variablen AUX1 bis AUX3. Diese Z-Werte werden nur für die QUICK CLUSTER-Prozedur verwendet. Weiterhin sortiert der erste Job die Fälle für eine Hot-Deck-ähnliche Ersetzung. Schließlich berechnet der Job noch die Anzahl der vollständigen Fälle auf den Variablen V1 bis V5 über die (Listwise-) Berechnung der Korrelationen:

```
* JOB1.
GET FILE='MISSING2.SYS'.
COUNT NMISS=  V1 TO V5 (MISSING).
COMPUTE SEQUENCE=$CASENUM.
SORT CASES BY NMISS.
DESCRIPTIVES AUX1 AUX2 AUX3 /OPTIONS 3.
CORRELATION V1 TO V5.
SAVE OUTFILE='SCRATCH'.
```

Die Anzahl vollständiger Beobachtungen auf den Variablen V1 to V5 ("6") muß manuell in das QUICK CLUSTER Kommando im zweiten Job eingefügt werden.

```
* JOB2.
GET FILE='SCRATCH'.
QUICK CLUSTER ZAUX1 TO ZAUX3
 /CRITERIA=CLUSTERS(6) NOUPDATE
 /INITIAL=FIRST
 /SAVE CLUSTER(CLUS).
SORT CASES BY CLUS NMISS.
```

```
IF (NMISS NE 0) V1=LAG(V1).
IF (NMISS NE 0) V2=LAG(V2).
IF (NMISS NE 0) V3=LAG(V3).
IF (NMISS NE 0) V4=LAG(V4).
IF (NMISS NE 0) V5=LAG(V5).
SORT CASES BY SEQUENCE.
```

QUICK CLUSTER sucht den nächsten (euklidische Distanz) Nachbarn auf den Variablen ZAUX1 bis ZAUX3. Die Clusterzugehörigkeit (CLUS) wird für jeden Fall gespeichert. Die Distanz jedes vollständigen Falles zum nächsten vollständigen Fall ist null. Das Sortieren nach CLUS und NMISS garantiert, daß ein vollständiger Fall zu Beginn des Datenfiles liegt. Fehlende Werte werden durch die Daten des nächsten (im Sinne der euklidischen Distanz) vollständigen Falles ersetzt. In diesem Job wurde ein variablenweises Hot-Deck verwendet, es kann aber durch die oben beschriebene fallweise Imputation ersetzt werden. Für einen Datensatz mit 1800 vollständigen Fällen benötigt ein 20Mhz-80386/10MHz-80287-System für diesen Job mehr als 70 Minuten.

Eine Ersetzungsmethode, die der von Dear (1959) vorgeschlagenen Verwendung der ersten Hauptkomponente sehr ähnelt, kann ebenso mit SPSS-PC+ realisiert werden. Da hierfür das FACTOR /SAVE Kommando verwendet werden muß, müssen die fehlenden Werte durch Mittelwerte ersetzt werden. Um die resultierenden Factorscores (die mit der Anderson/Rubin-Methode berechnet werden) wieder auf die ursprünglichen Einheiten zu bringen, werden die Mittelwerte und Standardabweichungen der Variablen benötigt. Diese können mit dem AGGREGATE-Kommando gespeichert werden. Der letzte Schritt besteht aus der Ersetzung durch die geschätzten Werte:

```
* DEAR-LIKE-SUBSTITUTION.
GET FILE='MISSING.SYS'.
* 1. REPLACE MD WITH MEANS (NECESSARY FOR
* FACTOR-SAVE COMMAND.
COMPUTE DUMMY=1.
SAVE OUTFILE='SCRATCH'.
AGGREGATE OUTFILE= *
/BREAK=DUMMY
/M1 TO M5 = MEAN (V1 TO V5)
/S1 TO S5 = SD (V1 TO V5).
* SAVE STANDARD DEVIATIONS FOR RESCALING OF 'Z-SCORES'.
JOIN MATCH /TABLE=* /FILE='SCRATCH'/BY DUMMY.
```

```
COMPUTE T1=V1.
COMPUTE T2=V2.
COMPUTE T3=V3.
COMPUTE T4=V4.
COMPUTE T5=V5.
IF (MISSING(V1)) T1=M1.
IF (MISSING(V2)) T2=M2.
IF (MISSING(V3)) T3=M3.
IF (MISSING(V4)) T4=M4.
IF (MISSING(V5)) T5=M5.
* 2. COMPUTE 1. PRINCIPAL COMPONENT.
FACTOR VARIABLES=T1 TO T5
 /ANALYSIS=ALL
 /CRITERIA=FACTORS(1)
 /EXTRACTION=PC
 /ROTATION=NOROTATE
 /SAVE=AR (1 DEAR).
* 3. RESCALE ESTIMATES.
COMPUTE T1=DEAR1*S1+M1.
COMPUTE T2=DEAR1*S2+M2.
COMPUTE T3=DEAR1*S3+M3.
COMPUTE T4=DEAR1*S4+M4.
COMPUTE T5=DEAR1*S5+M5.
* 4. REPLACE MISSING VALUES WITH ESTIMATES.
IF (MISSING(V1)) V1=T1.
IF (MISSING(V2)) V2=T2.
IF (MISSING(V3)) V3=T3.
IF (MISSING(V4)) V4=T4.
IF (MISSING(V5)) V5=T5.
```

Eine multiple Regressionsimputation ist mit SPSS-PC+ sehr einfach. Der einzige kleine Trick des Jobs besteht in der Beschränkung des Samples auf die vollständigen Fällen mit dem SELECT-Subkommando der Regressionsprozedur. Zusammen mit dem /MISSING= MEANSUBSTITUTION Subkommando zwingt dies SPSS-PC+ die Korrelationsmatrix "listwise" zu berechnen und die über Mittelwerte vervollständigten Daten für die Schätzung der vorhergesagten Werte zu verwenden:

```
GET FILE='MISSING.SYS'.
COUNT NMISS=V1 TO V5(MISSING).
REGRESSION VARIABLES=V1 TO V5
 /MISSING=MEANSUBSTITUTION
 /SELECT= NMISS EQ 0
```

```
  /DEPENDENT=V1/ENTER/SAVE PRED(M1)
  /DEPENDENT=V2/ENTER/SAVE PRED(M2)
  /DEPENDENT=V3/ENTER/SAVE PRED(M3)
  /DEPENDENT=V4/ENTER/SAVE PRED(M4)
  /DEPENDENT=V5/ENTER/SAVE PRED(M5).
IF (MISSING(V1)) V1=M1.
IF (MISSING(V2)) V2=M2.
IF (MISSING(V3)) V3=M3.
IF (MISSING(V4)) V4=M4.
IF (MISSING(V5)) V5=M5.
```

Eine iterierte multiple Regressionsimputationsmethode, die dem EM-Algorithmus für dieses Probiem entsprechen würde (mit der Ausnahme der zwar notwendigen, aber vernachlässigten Korrektur der Kovarianzmatrix) ist innerhalb von SPSS-PC+ ohne großen manuellen Aufwand nicht möglich. Dies würde getrennte multiple Regressionen für jede Menge von unvollständigen Beobachtungen mit einem anderen Muster der fehlenden Werte erfordern [12].

4. Die Leistungsfähigkeit der MD-Techniken

Selbstverständlich gilt das Interesse zunächst der Leistungsfähigkeit der MD-Techniken. Ihre Leistung hängt von sehr vielen Parametern ab, so z.B. dem fehlende Werte erzeugenden Mechanismus, dem Anteil fehlender Werte, der Anzahl der Variablen, der Anzahl der Fälle, dem Ausmaß der

[12] Eine einfache, aber sehr informative Analyse der fehlenden Werte ermöglicht ein "MD-pattern plot". Solch ein Plot ist in BMDPAM verfügbar. In SPSS-PC läßt sich dies so erreichen:

```
GET FILE='MISSING.SYS'.
COMPUTE PATTERN=0.
IF (MISSING(V1)) PATTERN=PATTERN+10000.
IF (MISSING(V2)) PATTERN=PATTERN+1000.
IF (MISSING(V3)) PATTERN=PATTERN+100.
IF (MISSING(V4)) PATTERN=PATTERN+10.
IF (MISSING(V5)) PATTERN=PATTERN+1.
RECODE V1 TO V5 (LOWEST THRU HIGHEST=1).
AGGREGATE OUTFILE=*
 /BREAK=PATTERN
 /NOFMDP=NU
 /M1 TO M5=LAST(V1 TO V5).
LIST PATTERN NOFMDP M1 TO M5.
```

Multikollinearität der Variablen, der Verteilungsform der Variablen und nicht zuletzt von den verwendeten Kriterien zur Beurteilung der Verfahren. Daher können nur für sehr einfache Methoden und spezielle Situationen (z.B. exakt ein fehlender Wert pro Variable) und nur unter der Annahme zufällig fehlender Werte analytische Aussagen über die Leistungsfähigkeit der Verfahren gemacht werden. Für den Fall systematisch fehlender Werte liegen keinerlei analytische Ergebnisse vor [13]. Der einzige Weg zur Beurteilung numerischer Verfahren, wenn keine analytischen Aussagen möglich sind, führt über Monte-Carlo-Simulationsstudien. Die MD-Literatur enthält zwar eine große Zahl von Simulationsstudien, von diesen werden aber nur ungefähr 10 häufiger zitiert [14]. Da es keinerlei Standardisierung in diesem Forschungsfeld gibt, können die meisten Studien nicht direkt verglichen werden. So verwenden die älteren Arbeiten nur selten ein experimentelles Design, die Ergebnisse können daher nicht auf bestimmte Faktorkombinationen zurückgeführt werden. Weiterhin verwenden die Arbeiten verschiedene Schätzmethoden und unterschiedliche Kriterien der Leistungsfähigkeit [15]. Trotzdem existiert in der Literatur weitgehend Konsens in bezug auf die prinzipielle Leistungsfähigkeit der Verfahren [16]. Die einfachen Ersetzungsverfahren (MEAN, CELL und Hot-Decks) werden in bezug auf die Schätzung von Kovarianzmatrizen als nicht so leistungsfähig betrachtet wie die multivariaten Verfahren. DEAR scheint nur bei hoher Multikollinearität brauchbar zu sein. Iterierte multiple Regressionen (insbesondere der EM-Algorithmus) werden als die

[13] Allerdings haben Santos (1981) für Regressionparameter und Brown (1983) für die Schätzung von Faktorladungen erste analytische Versuche unternommen.

[14] Die häufigst zitierten Arbeiten stammen von Haitovsky (1969), Timm (1970), Gleason/Staelin (1975), Beale/Little (1975), Kim/Curry (1977) und Little (1979). Die neueren Arbeiten von Vanguilder/Azen (1981), Kaiser (1984) und Basilevsky/ Sabourin/Hum/Anderson (1985) sind wesentlich sorgfältiger entworfen, mit nur wenigen Ausnahmen führen sie aber kaum zu anderen Ergebnissen.

[15] Eine interessante Konsequenz dieser Situation ist die Ablehnung der "pairwise"-Korrelationsschätzung, hauptsächlich auf der Basis der Simulationsergebnisse von Haitovsky (1969). Viele Simulationsstudien schlossen die "pairwise"-Schätzung daher schon im Design aus; mathematische Statistiker lehnten die Methode aus theoretischen Gründen ab (z.b. Heiberger 1977). Es ist bemerkenswert, daß in den letzten Jahren der Pairwise-Schätzer sowohl zunehmende theoretische Unterstützung (Brown 1983) und vor allem noch wesentlich bessere Simulationsergebnisse (Finkbeiner 1979, Basilevsky/Sabourin/Hum/Anderson 1985, Schnell 1985) erhalten hat als zuvor.

[16] Eine relativ aktuelle, wenn auch unvollständige Übersicht findet sich bei Anderson/Basilevsky/Hum (1983). Eine lesbare, wenn auch etwas veraltete Übersicht gibt Hamilton (1975). Eine kritische Übersicht über fast alle publizierten Arbeiten bis 1984 gibt Schnell (1986:138-157).

besten Verfahren angesehen. In Hinsicht auf die Schätzung einzelner fehlender Werte weichen die Ergebnisse kaum ab; dieses Problem wird aber in nur wenigen Arbeiten aufgegriffen. Die meisten Arbeiten konzentrieren sich auf die Schätzung von Summen und Mittelwerten.

Die wichtigste Beschränkung der Simulationsstudien liegt in der Verwendung nur eines fehlende Werte erzeugenden Prozesses: Mit der Ausnahme von Van Guilder/Azen (1981) basiert jede Simulationsstudie auf der Annahme zufällig fehlender Werte ("missing at random": MAR). Diese Annahme ist aber bei den meisten praktischen Problemen nicht erfüllt. Falls Empfehlungen auf der Basis der Ergebnisse von Simulationsstudien, die alle die MAR-Annahme verwenden, gegeben werden sollen, muß angenommen werden, daß es zwischen dem Ausfallprozeß und der Leistungsfähigkeit der Verfahren keinen Interaktionseffekt geben darf. Mit anderen Worten: Die Verfahren müssen unter jedem möglichen Ausfallmechanismus gleichmäßig leistungsfähig sein. Um diese kühne Annahme zu testen, wurde eine neue Simulationsstudie entworfen.

5. Entwurf der Simulationsstudie

Die Simulation basiert auf einem vollfaktoriellen Design mit vier Faktoren (Stichprobengröße, Interkorrelation, Prozentsatz fehlender Werte und fehlende Werte erzeugender Prozeß) mit jeweils 3 Stufen für jeden Faktor, so daß 3*3*3*3 = 81 verschiedene Bedingungen getestet wurden. Für jede experimentelle Bedingung wurden 10 verschiedene Datensätze mit jeweils 10 multivariatnormal verteilten Variablen mit gegebener Korrelationsstruktur erzeugt [17]. Jeder der so entstandenen 810 verschiedenen unvollständigen Datensätze wurde mit 10 verschiedenen MD-Techniken vervollständigt, so daß ein 3*3*3*3 vollfaktorielles Design mit wiederholter (*10) Messung auf dem letzten Faktor entstand.

Die Stichprobengröße wurde auf 100, 300 und 500 Fälle mit jeweils 10 Variablen festgelegt. Zusätzlich zu den 10 Variablen mit fehlenden Werten

[17] Das Simulationsprogramm wurde für den IBM-VS-FORTRAN Compiler (Level 1. 3.0) unter CMS geschrieben. Zur Erzeugung der Pseudozufallszahlen und für die statistischen Subroutinen wurden Subroutinen der Numerical Algorithm Group (NAG) Fortran-Library Mark 10 verwendet. Der grundlegende "multiplicative congruential generator" G05CAF wurde mit der Subroutine GO5CBF gestartet. Die multivariaten Normalverteilungen wurden mit den Generatoren GO5EAF und GO5EZF erzeugt. Die Simulation wurde auf der IBM 4341 der Universität Essen gerechnet und beanspruchte ca. 17 Stunden CPU-Zeit. Weitere Details und ein Programmlisting (ca. 700 Zeilen Fortran) finden sich bei Schnell (1986).

wurden drei Variablen ohne fehlende Werte erzeugt. Die Zusatzvariablen (V1 - V3) wurden als Klassifikationsvariablen und als unabhängige Variablen für die Regressionsmethoden verwendet. Die Zusatzvariablen sollten demographische Variablen wie Geschlecht und Alter simulieren, die in der Regel für jeden Fall eines Datensatzes vorliegen. Alle Berechnungen der Leistungsfähigkeit der Verfahren basieren auf den Schätzungen für die 10 unvollständigen Variablen und ihrer 45 Interkorrelationen. Die Struktur der Korrelationsmatrizen war homogen (in der Population): jedes Nicht-Diagonalelement der Korrelationsmatrix V4...V13 in der Population wurde gleich dem Wert des experimentellen Faktor CM (correlation mean) gesetzt. CM nahm die Werte 0.2, 0.4 und 0.6 an. Für die Simulation eines Thresholdmodelles, bei dem Werte dann fehlen, wenn eine ungemessene Variable einen stochastischen Schwellenwert überschreitet, wurde eine zusätzliche Hilfsvariable (V14) erzeugt, die nur zur Bestimmung der Ausfallwahrscheinlichkeit verwendet wurde. Die Korrelation dieser Variablen mit den Variablen V4 bis V1I3 wurde auf CM+O.1 festgesetzt. Die resultierende Korrelationsstruktur gibt Tabelle 1 wieder.

Tabelle 1: Gegebene Populationskorrelationsstruktur

	Hilfs-variablen			Haupt-variablen				Selektions-variable
	V1	V2	V3	V4	V5	...	V13	(V14)
V1		.2	.2	.15	.15	...	.15	.15
V2			.2	.15	.15	...	.15	.15
V3				.15	.15	...	.15	.15

Drei verschiedene MD erzeugende Prozesse wurde simuliert: Missing at Random (MAR), Missing at Random within Classes (MARC) und ein Hazard-Modell: Missing above Threshold (MAT).

Das Ausmaß fehlender Werte (MPER) für jeden Prozeß wurde auf 1%, 5% und 10% festgelegt. Die Löschung fehlender Werte erfolgte für alle Variablen unabhängig voneinander, so daß Fälle mit fehlenden Werten auf allen Hauptvariablen (V4 to V 13) möglich waren. Die Löschungen wurden im MAR-Modell zufällig vorgenommen. Das MARC-Modell basiert auf der Annahme, daß die Stichprobe aus zwei verschiedenen Populationen stammt, die sich in Hinsicht auf die Ausfallwahrscheinlichkeiten und die Interkorrelationsstruktur unterscheiden. Innerhalb der Populationen fehlen die Daten zufällig. In der Simulation wurden die erzeugten Stichproben mit einer Wahrscheinlichkeit von 0.8 aus einer Subpopulation mit einer Interkorrelation von

CM und mit der Wahrscheinlichkeit 0.2 aus einer Subpopulation mit einer Interkorrelation von 0.1 gezogen. Das Verhältnis der Ausfallwahrscheinlichkeiten wurde so auf 1:10 festgelegt, daß der Anteil fehlender Werte im kombinierten Sample dem Prozentsatz der experimentellen Bedingung (MPER) entsprach. Der dritte simulierte Prozeß war ein Hazard-Modell. Falls für einen gegebenen Fall eine ungemessene Variable einen unbekannten Schwellenwert überschreitet, fällt der Fall aus. Dieses MAT-Modell wurde implementiert, indem eine einfache Zufallstichprobe erzeugt wurde, bei der dann diejenigen Fälle gelöscht wurden, bei denen eine Selektionsvariable (V14) einen Schwellenwert überschritt. Für jeden der drei Level von MPER wurde ein Schwellenwert gewählt, der zu dem gewünschten Ausmaß fehlender Werte führte.

Zehn verschiedene MD-Techniken (PROG) wurden verwendet:

1. - LISTWISE (vollständige Fälle)
2. - PAIRWISE (paarweise vorhandene Werte)
3. - MEAN (Mittelwertersetzung)
4. - CELLMEAN (Zellenmittelwertersetzung)
5. - HOT-DECK (Sequentielle variablenweise Fallverdoppelung)
6. - REGRESSION (Einfache Regression)
7. - HAZARD (Regression mit MD-Indikator)
8. - DEAR (Hauptkomponente)
9. - MULT (Multiple Regression)
10. - ILS (iterierte multiple Regression)

Für jede unvollständige Variable verwendete CELLMEAN den Zellenmittelwert der vorhandenen Daten einer Crossbreak-Tabelle der Variable mit den ersten beiden zusätzlichen Variablen (V1,V2). HOT-DECK verwendete eine Kreuztabelle der Fälle mit den Variablen V1 und V2 als Imputationsklassen. REGRESSION verwendete die am stärksten korrelierende Variable als Prädiktor. HAZARD verwendete neben derselben Variable wie REGRESSION die Anzahl der fehlenden Werte des gegebenen Falles als Prädiktor. DEAR ersetzte die fehlenden Werte durch Schätzungen auf der Basis der ersten Hauptkomponente, die aus der PAIRWISE-Korrelationsmatrix berechnet wurde. MULT verwendete eine multiple Regression mit allen verfügbaren Variablen eines gegebenen Falles (die Prozedur entspricht damit der von Buck, allerdings ohne Korrektur der Kovarianzmatrix). Die resultierenden Schätzungen von MULT bildeten den Ausgangspunkt für ILS.

ILS wiederholte die multiplen Regressionen bis kein Korrelationskoeffizient in zwei aufeinanderfolgenden Schritten sich um mehr als 0.0001 vom vorherigen unterschied.

Zwei Aspekte der Leistungsfähigkeit der MD-Techniken sollten untersucht werden: Die Güte der Schätzung von Korrelationskoeffizienten und die Güte der Schätzung einzelner fehlender Werte.

Der Bias der geschätzten Korrelationsmatrizen wurde berechnet als die mittlere Abweichung der geschätzten Korrelationen von den tatsächlichen Korrelationen:

$$RDEV = \frac{\sum_{i=1}^{p-1} \sum_{j=i+1}^{p} RT_{i,j} - R_{i,j}}{p\,(1-p)}$$

Timm (1970:426) verwendete die euklidische Norm der Differenzenmatrix: Tr [(RT-R) (RT-R)']; in Anlehnung an Gleason/Staelin (1975:242) wurde eine Modifikation dieses Maßes, das "root mean square residual [18]", in der Simulation verwendet:

$$WRSDEV = \left[\frac{\sum_{i=1}^{p-1} \sum_{j=i+1}^{p} (RT_{i,j} - R_{i,j})^2}{p\,(1-p)} \right]^{1/2}$$

Als drittes Ähnlichkeitsmaß für die Matrizen wurde der Korrelationskoeffizient der beiden als Datenvektoren aufgefaßten Korrelationsmatrizen berechnet [19]. Dieser Pearson-Korrelationskoeffizient wurde dann mit der Fisher-Transformation in z-Werte umgerechnet:

$$ZMKS = 0.5 \ln [(1+r)/(1-r)]$$

Die Leistungsfähigkeit der Ersetzungsverfahren (DECK, REGRESSION, HAZARD, MULT, ILS) [20] in Hinsicht auf die Schätzung [21] einzelner fehlender Werte wurde durch vier verschiedene Maße erfaßt.

[18] Dies ist mit dem "overall deviation index" von Kim/Curry (1977: 224) identisch.

[19] In GAUSS ist dies mit wenigen Tastendrücken möglich: R=CORRX (VEC (RT) ~VEC(R)).

[20] Obwohl MEAN und CELLMEAN fehlende Werte ersetzen, wurden sie als triviale Ersetzungsmethoden aufgefaßt, wenn man an der Schätzung einzelner fehlender Werte interessiert ist.

Das erste Maß ist die "root mean square deviation" (z.B. von Kalton/ Santos 1983 verwendet):

$$ERSD = \left[\frac{\Sigma \Sigma (DT_{i,j} - D_{i,j})^2}{M} \right]^{1/2}$$

wobei DT die wahre Datenmatrix, D die geschätzte Datenmatrix und M die Anzahl fehlender Werte darstellt. Die mittlere Abweichung wurde berechnet als:

$$EMD = (\Sigma DT_{i,j} - D_{i,j}) / M,$$

die mittlere absolute Abweichung als

$$EMAD = \Sigma |DT_{i,j} - D_{i,j}| / M.$$

Die letzte Maßzahl ist die Anzahl der exakt geschätzten fehlenden Werte (geschätzter Wert=wahrer Wert), dividiert durch die Anzahl der fehlenden Werte. Dieses Maß wird als HIT bezeichnet.

Ergebnisse der Simulation

Aufgrund des beschränkten Raumes werden die wichtigsten Ergebnisse als Plots präsentiert [22]. Die Y-Achse gibt den Mittelwert der abhängigen Variable über die 10 Wiederholungen unter denselben experimentellen Bedingungen für eine gegebene MD-Technik wieder. Die X-Achse (experimentelle Bedingung) wurde hier berechnet als COND= (MTYPE-1)*27 + (MPER-1)*9 + (CM-1)*3 + N. Daher entspricht COND=1 der Bedingung MAR (MTYPE=1), 1% MD (MPER=1), mittlere Interkorrelation gleich 0.2 (CM=1) und Stichprobengröße gleich 100 (N=1). Folglich wechselt in den Plots der fehlende Werte erzeugende Prozeß alle 27 Einheiten, MPER alle 9 Einheiten, CM alle 3 Einheiten und N mit jeder Einheit. Um einen bildlichen Eindruck der Leistung der Verfahren zu geben, wurden die jeweils 81 Meßpunkte einer MD-Technik mit einer Linie zu einem Profil über die experimentellen Bedingungen verbunden.

[21] Mit der Ausnahme des Hot-Decks erzeugen alle Verfahren keine ganzzahligen Schätzungen. Da die Eingabewerte ganzzahlig waren (nach der Erzeugung kategorisiert in 7 Kategorien), wurden für die Berechnung der Datenähnlichkeitsmaße die Schätzungen kategorisert (FORTRAN: DNINT).

[22] Tabellen und multivariate Analysen (MANOVA's, Regressionen, Clusteranalysen) für eine Reihe spezieller Fragen finden sich bei Schnell (1985, 1986). Der Datensatz (SPSS-PC-Systemfiles und ASCII-Version) wird auf Anfrage zur Verfügung gestellt.

Plot 1 zeigt die mittlere Abweichung des geschätzten Korrelationskoeffizienten für die 81 experimentellen Bedingungen. Der Einfluß der experimentellen Bedingungen ist deutlich sichtbar. Obwohl die Profile ähnlich erscheinen, gibt es deutliche Unterschiede zwischen den Verfahren. Plot 2 zeigt dies am Beispiel der Profile für LISTWISE und PAIRWISE. Unter nahezu jeder getesteten Bedingung zeigt LISTWISE schlechtere Resultate. Die Differenzen zwischen LISTWISE und PAIRWISE erreichen bis zu .13 unter denselben experimentellen Bedingungen.

Auf den ersten Blick erscheinen die mittleren Abweichungen nicht sehr groß: Der Mittelwert unter MAR erreicht 0.03, unter MAT 0.05 und -0.04 innerhalb von MARC selbst für die schlechteste Technik (ein positives RDEV impliziert eine Unterschätzung). Für LISTWISE wird das absolute Maximum von RDEV mit 0.15 innerhalb von MAR, 0.18 innerhalb von MAT und -0.17 innerhalb von MARC erreicht. Für ILS beträgt die maximale Abweichung -0.06 innerhalb von MARC. Betrachtet man aber die "Struktur" der Korrelationsmatrix (ZMKS: Plot 3), so zeigt sich, daß scheinbar kleine Abweichungen einzelner Koeffizienten zu großen Abweichungen bei der Struktur der Korrelationsmatrix führen können: Selbst der Mittelwert von ZMKS erreicht Minimumwerte zwischen 0.58 (LISTWISE) und 1.16 (ILS). Selbstverständlich zeigen die Minimumwerte für ZMKS unter den einzelnen Bedingungen noch extremere Werte. So erreicht z.B. innerhalb von MAR LISTWISE den Wert 0.3 (r=0.29) als sein Minimum, dagegen ILS 1.22 (r=.84). Innerhalb von MARC erreicht ILS sein Minimum mit .76 (r=.64). Das Ergebnis zeigt, daß selbst eine der besten Techniken zu irreführenden Ergebnissen führen kann.

Unter praktischen Gesichtspunkten ist der Anteil korrekt geschätzter fehlender Werte (HIT) von Bedeutung (Plot 4). Trotz der Kategorisierung liegt der Anteil korrekt geschätzter fehlender Werte für die meisten Verfahren zwischen 0.25 and 0.40. Eines der besten Verfahren, ILS, erreicht im Mittel 38% innerhalb von MAR, 36% innerhalb von MAT und 33% innerhalb von MARC. HIT erreicht bei hohem CM und großen Stichproben relativ unabhängig von der Anzahl fehlender Werte Maximumwerte von 45%. Aber Plot 4 zeigt bemerkenswerte Unterschiede zwischen den Verfahren. Insbesondere DEAR ist gegenüber dem erzeugenden Prozeß hochempfindlich. Das Hot-Deck ist bei weitem das schlechteste Verfahren. MULT und ILS sind dagegen kaum unterscheidbar (Plot 4a).

Falls man an der Schätzung einzelner fehlender Werte interessiert ist, so ist die mittlere Abweichung der geschätzten Werte von den wahren Werten

von Interesse. Plot 5 zeigt eine starke Abhängigkeit von EMD vom fehlende Werte erzeugenden Prozeß. Die Unterschätzung innerhalb von MAR zwischen 0.017 und 0.025 steigt auf 0.295 und 1.05 innerhalb von MAT. Die Unterschiede zwischen den Verfahren sind deutlich sichtbar, insbesondere die schlechte Leistung des Hot-Decks [23].

Obwohl der Einfluß der experimentellen Faktoren auf die Leistungsfähigkeit der MD-Techniken sowohl von der Technik als auch vom fehlende Werte erzeugenden Prozeß abhängen [24], sind ein paar Tendenzen in den Plots (und den hier nicht berichteten Regressionen) evident. Da die Anzahl der Fälle fast niemals einen Effekt besitzt, scheint es angebracht, in weiteren Simulationen auf diesen Faktor zu verzichten. Die Leistung der Verfahren in Hinsicht auf die Schätzung einzelner fehlender Werte (EMD, EMAD, ERSD, HIT) kann nur teilweise durch die experimentellen Faktoren erklärt werden, das R^2 der multiplen Regressionen überschreitet selten 0.30 für HIT und 0.60 für ERSD. Die relative Wichtigkeit von MPER gegenüber CM hängt von einer Reihe von Faktoren ab, im allgemeinen ist CM aber für die multivariaten Verfahren (DEAR, MULT, ILS) wichtiger als für die anderen Techniken. Je mehr Redundanz in den Daten steckt, umso besser sind die Schätzungen. Wie die Plots (und die hier nicht berichteten MANOVA's) zeigen, kann nahezu jeder mögliche Interaktionseffekt gefunden werden, so daß weitere Generalisierungen kaum möglich sind. Für die multiplen Regressionen für die Maße der Güte der Schätzungen der Korrelationen (RDEV, WRSDEV, ZMKS) kann ein weit besserer Fit festgestellt werden (R^2 zwischen 0.75 and 0.87). In diesen Regressionen ist der Einfluß der Anzahl der MD nahezu immer stärker als der Effekt der Stärke der Interkorrelation. "Typische" Beta-Werte für CM liegen zwischen -0.10 und -0.20, für MPER zwischen -0.75 and -0.90.

Zusammenfassung

Die Ergebnisse zeigen die starke Abhängigkeit aller Verfahren vom - in der Regel unbekannten - fehlende Werte erzeugenden Prozeß. Das wichtigste Resultat der Simulation ist daher die Feststellung, daß ein "bestes" Schätzverfahren nicht existiert. Selbst die "besten" Verfahren brechen unter

[23] Dies Ergebnis stimmt mit den Resultaten von Kaiser (1983) überein.

[24] Alle MANOVA-Interaktionsterme des Typs MTYPE*PROG, MTYPE*N, MTYPE*CM, MTYPE*MPER und die meisten der PROG*N, PROG*CM, PROG*MPER Interaktionen sind signifikant mit $p < 0.001$. Die meisten der entsprechenden 3-Wege-Interaktionen sind ebenfalls signifikant.

bestimmten Bedingungen zusammen. Die weitere Suche nach "besten" Schätzverfahren erscheint daher sinnlos. Die Ergebnisse der Simulation legen nahe, daß ein angemessener Umgang mit fehlenden Werten nicht darin bestehen kann, sie zu vernachlässigen oder sie mit einer "automatischen" Prozedur wie BMDPAM zu beseitigen, sondern nur darin, unterschiedliche plausible theoretische Annahmen über den fehlende Werte erzeugenden Prozeß während der Imputation zu simulieren. Solche "multiplen Imputationen" (Rubin 1987) scheinen der einzige Weg bei der Analyse eines Datensatzes mit einem nicht-trivialen Anteil fehlender Werte zu sein.

Literatur

ANDERSON, A.B./BASILEVSKY, A./HUM, D.J. (1983): Missing Data: A Review of the Literature; in: ROSSI, P.H./ WRIGHT, J.D./ ANDERSON, A.B. (eds.): Handbook of Survey Research, New York, p.415-493.

BASILEVSKY, A./SABOURIN, D./HUM, D./ANDERSON, A. (1985): Missing Data Estimators in the General Linear Model: An Evaluation of Simulated Data as an Experimental Design; in: Communications in Statistics, Simulation and Computation, 14, 2, p.371 -394.

BEALE, E.M.L./LITTLE, R.J.A. (1975): Missing Values in Multivariate Analysis; in: Journal of the Royal Statistical Society, Series B, 37, p. 129-146.

BERGER,M.P.F. (1979): A FORTRAN IV Program for the Estimation of Missing Data; in: Behavior Research Methods and Instrumentation, 11, 3, p.395-396.

BERK, R.A./RAY, S.C. (1982): Selection Biases in Sociological Data; in: Social Science Research, 11, p.352-398.

BERK, R.A. (1983): An Introduction to Sample Selection Bias in Sociological Data; in: American Sociological Review, 48, p.386-398.

BROWN, C.H. (1983): Asymptotic Comparison of Missing Data Procedures for Estimating Factor Loadings; in: Psychometrika, 48, 2, p.269-291.

BUCK, S.F. (1960): A Method of Estimation of Missing Values of Multivariate Data Suitable for Use With an Electronic Computer; in: Journal of the Royal Statistical Society, Series B, 22, p.302-307.

CHAPMAN, D.W. (1983): An Investigation of Nonresponse Imputation Procedures for the Health and Nutrition Examination Survey; in: MADOW, W.G./NISSELSON, H./OLKIN, I. (eds.): Incomplete Data in Sample Surveys, Vol.l, p.435-483, New York.

COHEN, J./COHEN, P. (1975): Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, Hillsdale.

DEAR, R.E. (1959): A Principal Component Missing Data Method for Multiple Regression Models. Technical Report SP-86, System Development Corporation, Santa Monica, Calif.

DEMPSTER, A.P./LAIRD, N.M./RUBIN, D.B. (1977): Maximum Likelihood from Incomplete Data via the EM-Algorithm; in: Journal of the Royal Statistical Society, Series B, 3g, p.1-22.

DILLON, W.R./GOLDSTEIN, M. (1984): Muitivariate Analysis. Methods and Applications, New York.

FAY, R.E. (1986): Causal Models for Patterns of Nonresponse; in: JASA, 81, 394, S.354-365.

FEDERSPIEL, C.F./MONROE, R.J./GREENBERG, B.G. (1959): An Investigation of Some Multiple Regression Methods for Incomplete Samples, University of North Carolina Institute of Statistics, Mimeo Series #236.

FIELDING, A. (1979): Binary Segmentation: The Automatic Interaction Detector and Related Techniques for Exploring Data Structure; in: O'MUIRCHEARTAIGH, C.A./PAYNE, C. (eds.): The Analysis of Survey Data, Vol.l, p.221 -257, Chichester.

FINKBEINER, C. (1979): Estimation for the Multiple Factor Model when Data are Missing; in: Psychometrika, 44, p.409-420.

FORD, B.L. (1983): An Overview of Hot-Deck Procedures; in: MADOW, W.G./ OLKIN, I. / RUBIN, D.B. (eds.): Incomplete Data in Sample Surveys, New York, Vol. 2, p.l 85-207.

FRANE, J. (1978): Missing Data and BMDP: Some Pragmatic Approaches; in: ASA Proceedings of the Statistical Computing Section, p.27-33.

FUCHS, C. (1982): Maximum Likelihood Estimation and Model Selection in Contingency Tables With Misssing Data; in: JASA, 77, 378, S.270-278.

GLASSER, M. (1964): Linear Regression Analysis With Missing Observations Among the Independent Variables; in: JASA, 59, p. 834-844.

GLEASON, T.C./STAELIN, R. (1975): A Proposal for Handling Missing Data; in: Psychometrika, 40, p.229-252.

GREENE, W.H. (1986): LIMDEP User's Manual.

HAAGENAARS, J.A. (1988): Log-Linear Analysis with Latent Variables and Missing Data; Paper presented at the International Conference on Social Science Methodology, Dubrovnik, May/June 1988.

HAITOVSKY, Y. (1968): Missing Data in Regression Analysis; in: Journal of the Royal Statistical Society, Series B, 63, p.67-82.

HAMILTON, M.A. (1975): Regression Analysis When There are Missing Observations - A Survey and Bibliography. Technical Report 1-3-75. Statistical Laboratory, Montana State University.

HARTIGAN, J.A. (1975): Clustering Algorithms, New York.

HECKMAN, J.J. (1976): The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dedendent Variables and a Simple Estimator for such Models; in: Annals of Economic and Social Measurement, 5/4, p.475-492.

HEIBERGER, R.M. (1977): Regression with the Pairwise-Present Covariance Matrix: A Dangerous Practice; in: ASA Proceedings of the Statistical Computing Section, p. 38-47.

HILL, M./DIXON, W.J. (1981): Missing Data: Search for Patterns; in: ASA Proceedings of the Statistical Computing Section, p.57-60.

KAISER, J. (1983): The Effectiveness of Hot-Deck Procedures in Small Samples; in: ASA Proceedings of the Section on Survey Research Methods, p.523-528.

KAISER, J. (1984): The Estimation of Missing Values, Ph.D. Thesis, University of Kansas.

KALTON, G./KASPRZYK, D. (1982): Imputing for Missing Survey Responses; in: ASA Proceedings of the Section on Survey Research Methods, p.22-31.

KALTON, G. (1983): Compensating for Missing Survey Data, Survey Research Center, Institute for Social Research, University of Michigan.

KIM, J.O./CURRY, J. (1977): The Treatment of Missing Data in Multivariate Analysis; in: Sociological Methods and Research, 6, 2, p. 215-239.

LITTLE, R.J.A. (1979): Maximum Likelihood Inference for Multiple Regression with Missing Values: A Simulation Study; in: Journal of the Royal Statistical Society, Series B, 41, 1, p.76-87.

LITTLE, R.J.A. (1983): The Nonignorable Case; in: MADOW,W.G./ OLKIN, I./RUBIN, D.B. (eds.): Incomplete Data in Sample Surveys, Vol 2, p.383-413, New York.

LITTLE, R.J.A./RUBIN, D.B. (1987): Statistical Analysis With Missing Data, New York (Wiley).

MÖNTMANN, V./BOL LINGER, G./HERRMANN, A. (1983): Tests auf Zufälligkeit von "Missing Data"; in: WILKE,H.(ed.): Statistik Software in der Sozialforschung, p.87-101, Berlin.

NELSON, F.D. (1977): Censored Regression Models with Unobserved, Stochastic Censoring Thresholds; in: Journal of Econometrics, 6, 1977, p.309-327.

O'GRADY, K.E. (1982): Regression Estimation of Missing Data; in: Behavior Research Methods and Instrumentation, 14, 3, p.359-360.

OH, H.L./SCHEUREN, F.J. (1983): Weighting Adjustment for Unit Nonresponse; in: MADOW,W.G./OLKIN,I./RUBIN,D.B. (eds.): Incomplete Data in Sample Surveys, New York, Vol. 2, p.143-184.

ORCHARD, T./WOODBURY, M.A. (1970): A Missing Information Principle: Theory and Practice; in: Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, 1, p.697-715.

RIZVI, H. (1983): An Empirical Investigation of Some Item Nonresponse Adjustment Procedures; in: MADOW,W.G./ NISSELSON,H./ OLKIN, I. (eds.): Incomplete Data in Sample Surveys, Vol.l, p.299-366, New York.

RUBIN, D.B. (1976): Inference and Missing Data; in: Biometrika, 63, p.581 - 592.

RUBIN, D.B. (1987): Multiple Imputation for Nonresponse in Surveys, New York (Wiley).

SANDE, G. (1979): Numerical Edit and Imputation; in: Bulletin of the International Statistical Institute, 42nd Session of the International Statistical Institute, p.455-463.

SANDE, I.G. (1982): Imputation in Surveys: Coping With Reality; in: The American Statistician, 36, 3, 1982, p.145-152.

SANTOS, R.L. (1981): Effects of Imputation on Compiex Statistics, Survey Research Center, Institute for Social Research, University of Michigan.

SCHMEE, J./HAHN, G.J. (1979): A Simple Method for Regression Analysis with Censored Data; in: Technometrics, 21, 4, p.417-432.

SCHNELL, R. (1985): Zur Effizienz einiger Missing-Data-Techniken; in: ZUMA Nachrichten 17, p.50-74.

SCHNELL, R. (1986): Missing-Data-Probleme in der empirischen Sozialforschung, Dissertation, Bochum.

SONQUIST, J.A./BAKER, E.L./MORGAN, J.N. (1971): Searching for Structure, Institute for Social Research, University of Michigan, Ann Arbor.

TIMM, N.H. (1969): Estimating Variance-Covariance and Correlation Matrices from Incomplete Data, University of California, Berkeley, Ph.D.Thesis.

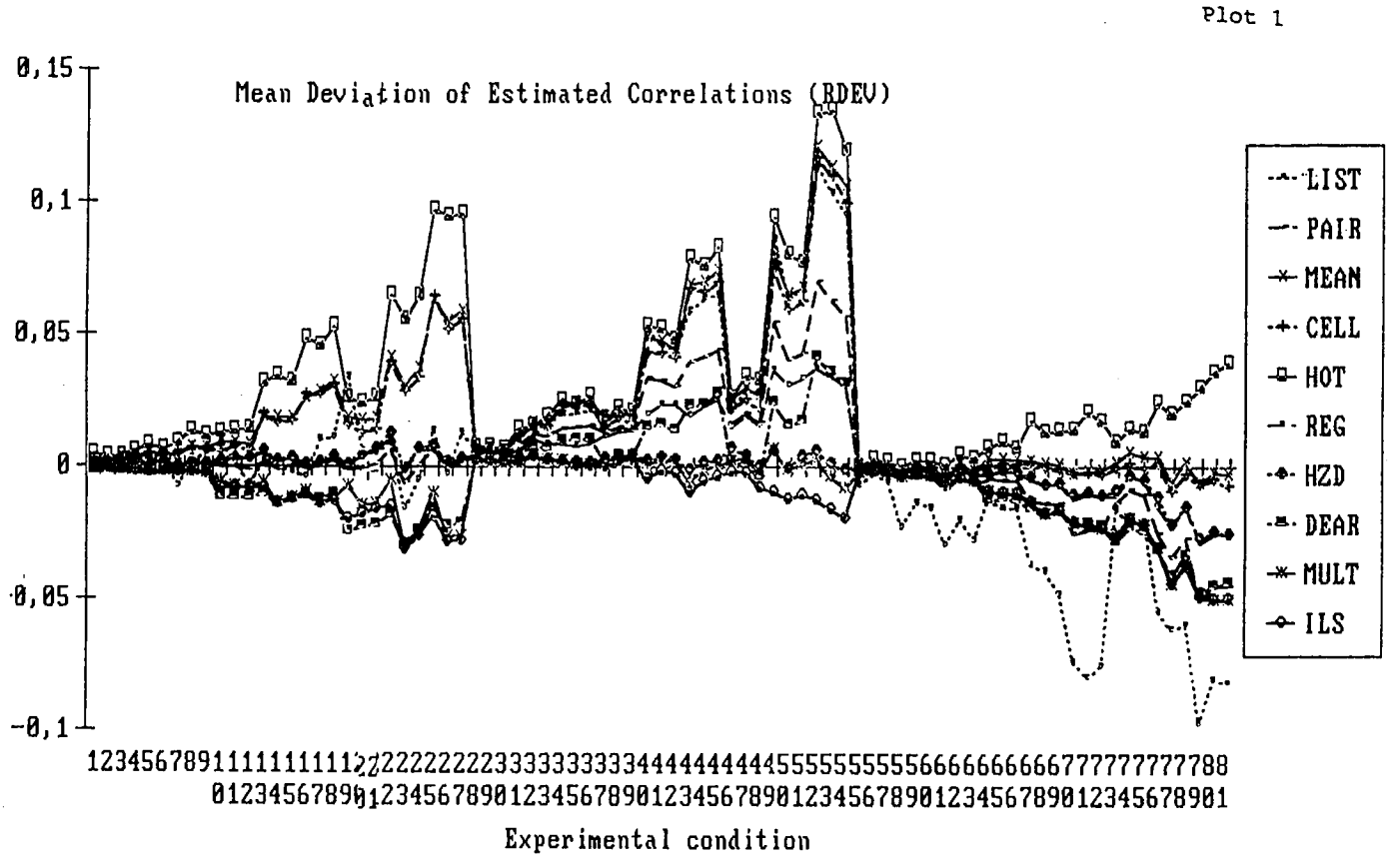
TIMM, N.H. (1970): The Estimation of Variance-Covariance and Correlation Matrices from Incomplete Data; in: Psychometrika, 35, p.417-437.

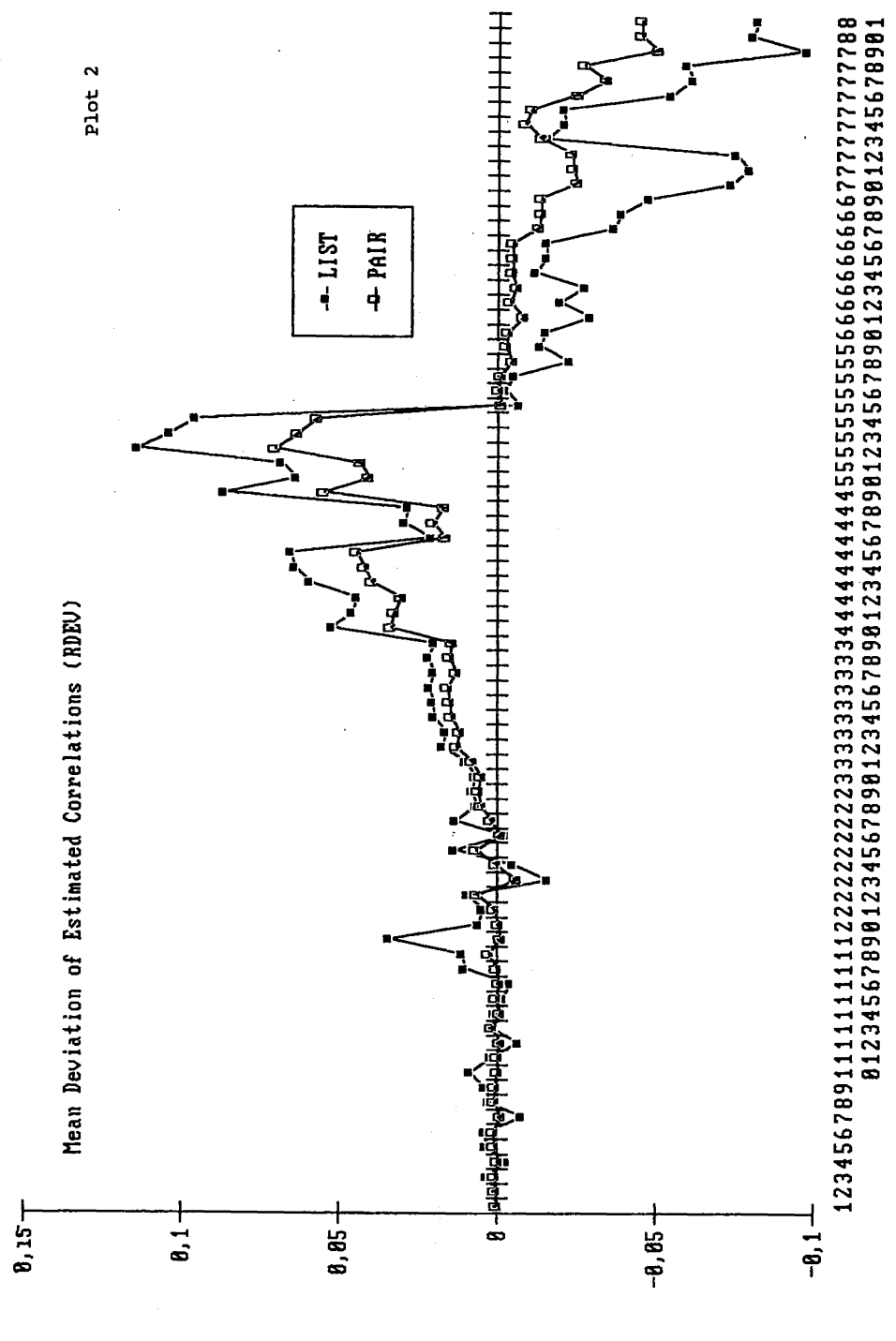
VACEK, P.M./ASHIKAGA, T. (1980): An Examination of the Nearest Neighbor Rule for Imputing Missing Values; in: ASA Proceedings of the Statistical Computing Section, p.326-331.

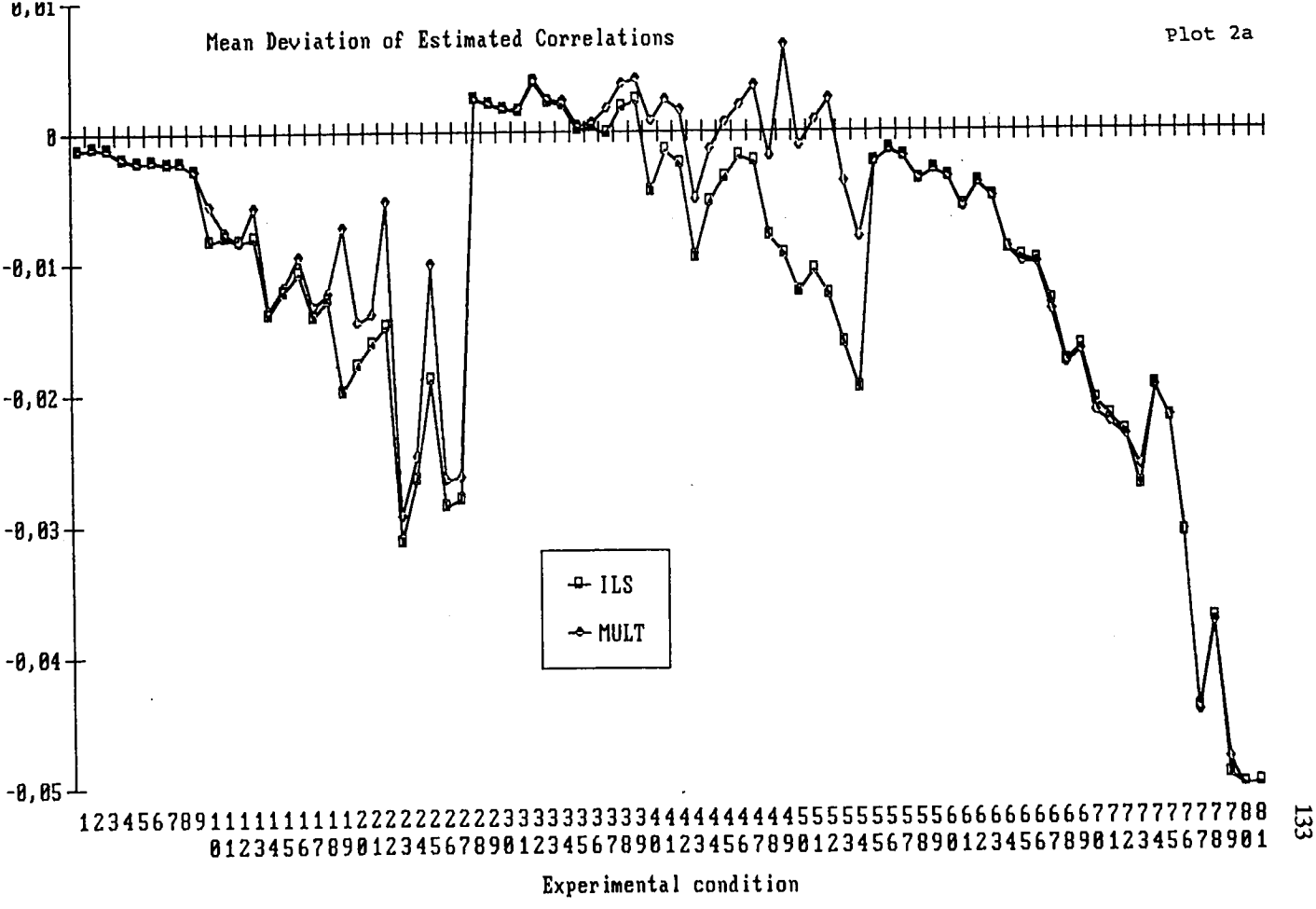
VAN GUILDER, M./AZEN, S. (1981): Conclusions Regarding Algorithms for Handling Incomplete Data; in: ASA Proceedings of the Statistical Computing Section, p.53-56.

WISHART, D. (1978): Treatment of Missing Values in Cluster Analysis; in: International Association for Statistical Computing: COMPSTAT 1978, p.281 -287, Wien.

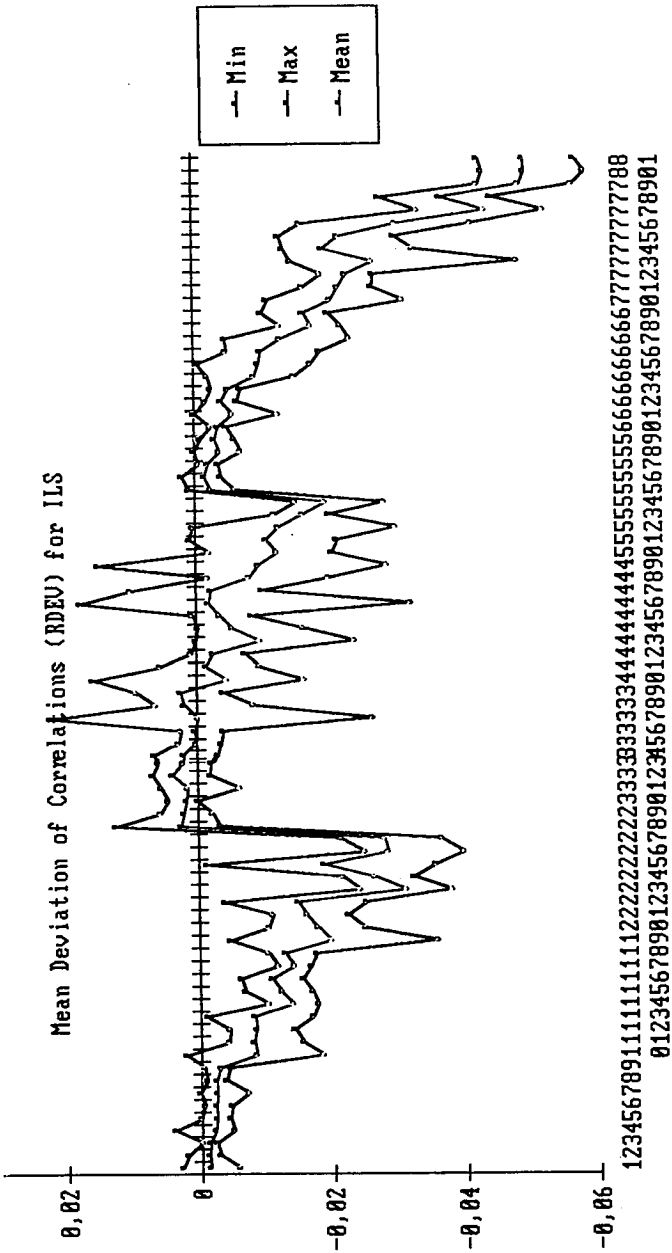
WISHART, D. (1985): Estimation of Missing Values and Diagnosis Using Hierarchical Classifications; in: Computational Statistics Quarterly, 2, 1, 1985, p. 25-134.

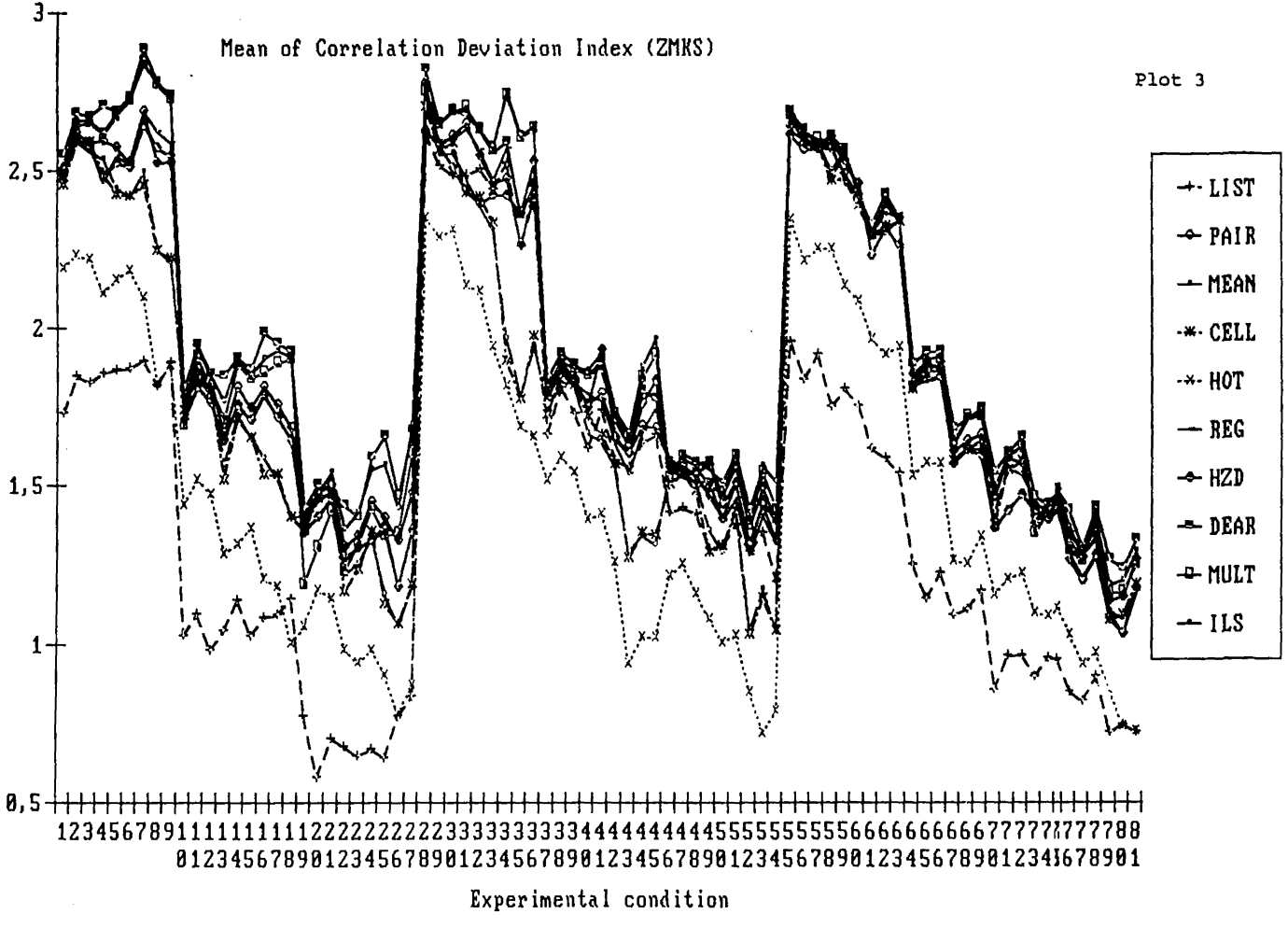

Plot 1
Mean Deviation of Estimated Correlations (RDEV)
0,15
0,1
0,05
0
-0,05
-0,1
LIST
PAIR
MEAN
CELL
HOT
REG
HZD
DEAR
MULT
ILS
Experimental condition

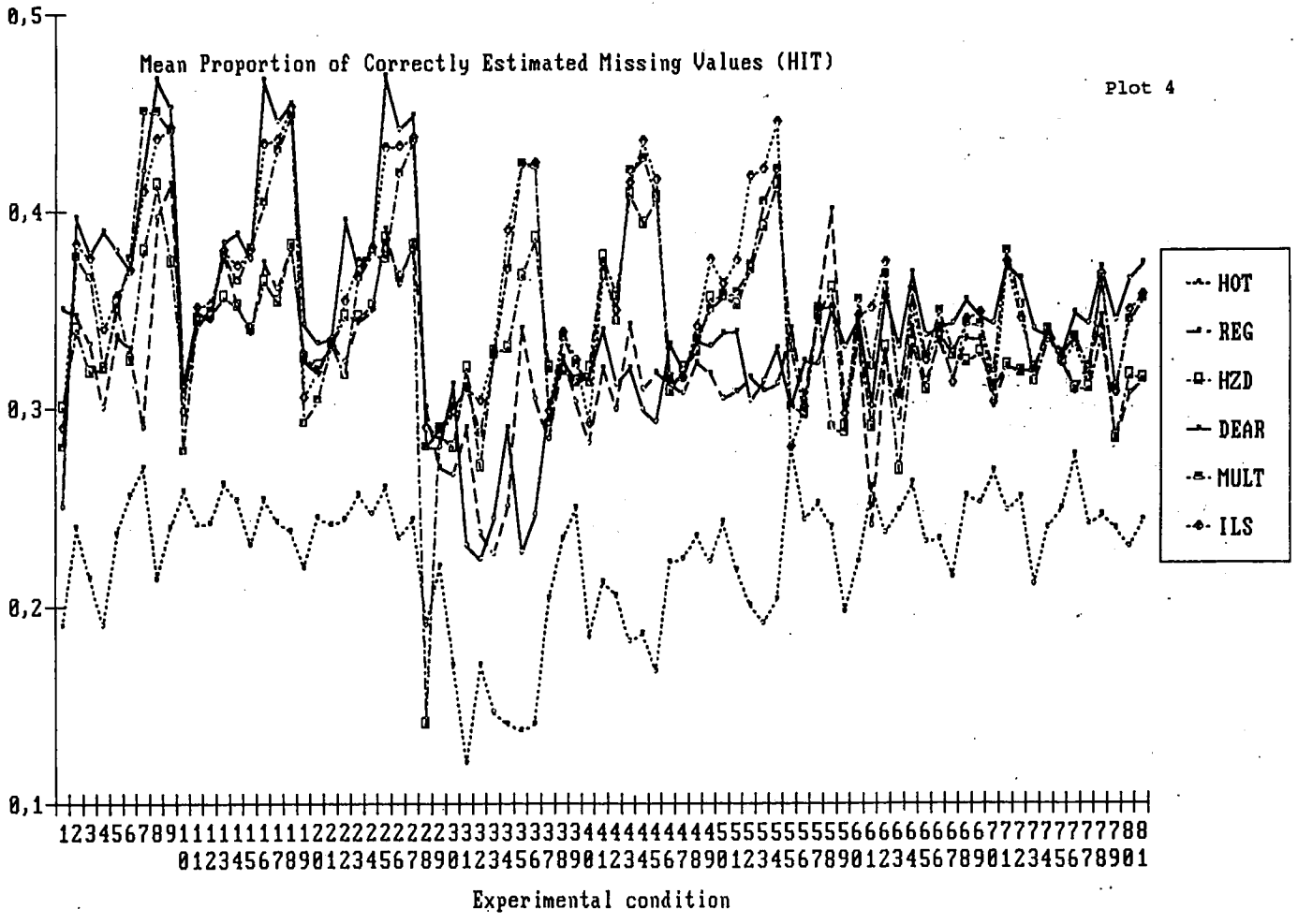

Plot 2
Mean Deviation of Estimated Correlations (RDEV)
LIST
PAIR
0,15
0,1
0,05
0
-0,05
-0,1
Experimental condition

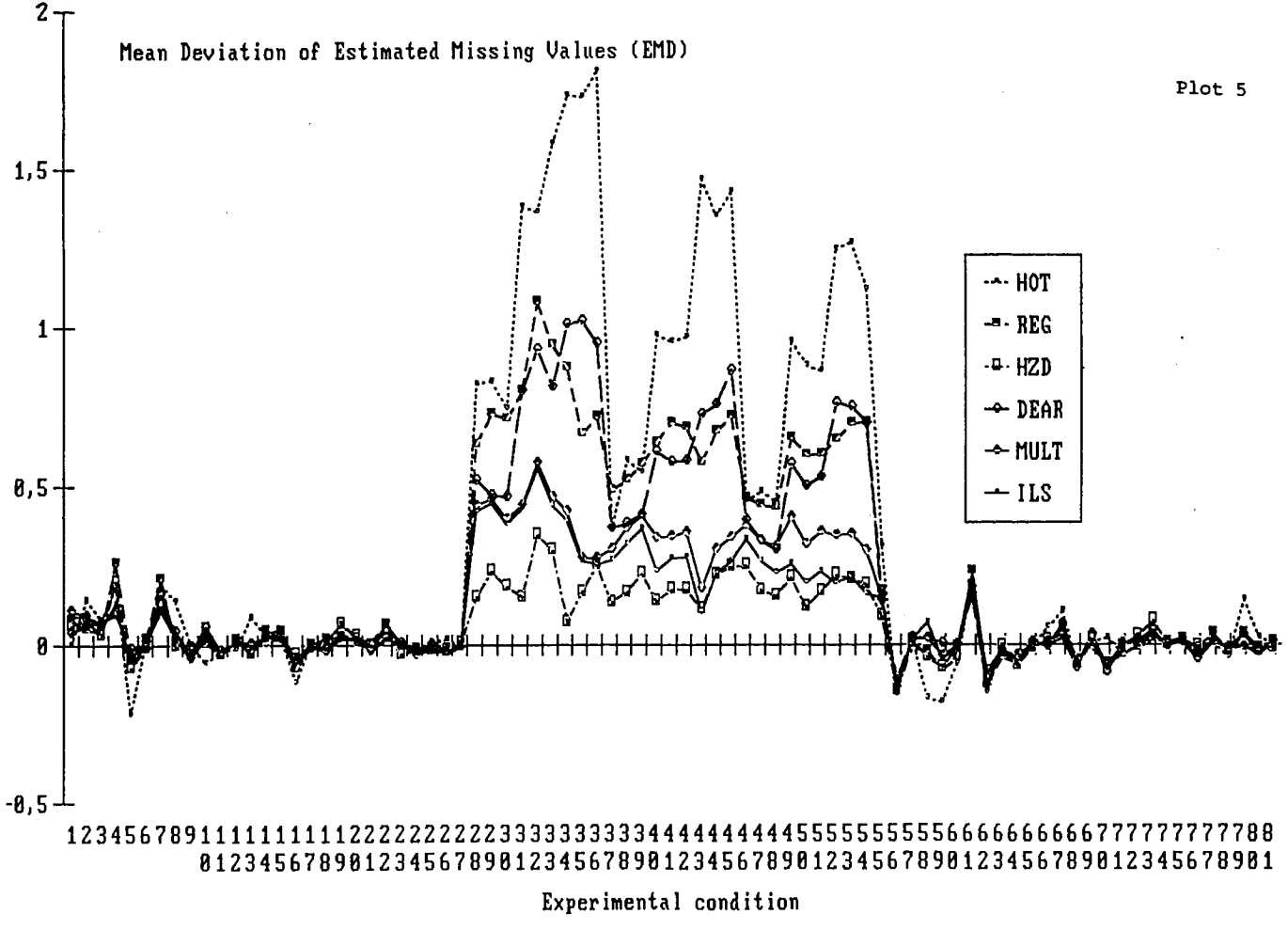

Mean Deviation of Estimated Correlations
Plot 2a
0,01
0
-0,01
-0,02
-0,03
-0,04
-0,05
ILS
MULT
1234567891111111111122222222223333333333444444444455555555556666666666677777777788
0123456789012345678901234567890123456789012345678901234567890123456789012345678901
Experimental condition

Plot 2b


Mean Deviation of Correlations (RDEV) for ILS
Min
Max
Mean
0,02
0
-0,02
-0,04
-0,06
Experimental Condition


Mean of Correlation Deviation Index (ZMKS)
Plot 3
LIST
PAIR
MEAN
CELL
HOT
REG
HZD
DEAR
MULT
ILS
3
2,5
2
1,5
1
0,5
Experimental condition


Mean Proportion of Correctly Estimated Missing Values (HIT)
Plot 4
0,5
0,4
0,3
0,2
0,1
HOT
REG
HZD
DEAR
MULT
ILS
Experimental condition


Mean Deviation of Estimated Missing Values (EMD)
Plot 5
HOT
REG
HZD
DEAR
MULT
ILS
2
1,5
1
0,5
0
-0,5
Experimental condition